Bike sharing systems: a new incentive rebalancing method based on spatial outliers detection

Yousra Chabchoub, Rayane El Sibai, Christine Fricker

Bike sharing systems: a new incentive rebalancing method based on spatial outliers detection

Yousra Chabchoub

LISITE Laboratory, Institut suprieur d'lectronique de Paris (ISEP)
10 Rue de Vanves, Issy-les-Moulineaux 92130, France
E-mail: yousra.chabchoub@isep.fr

Rayane El Sibai

LISITE Laboratory, Institut suprieur d'lectronique de Paris (ISEP)
10 Rue de Vanves, Issy-les-Moulineaux 92130, France
E-mail: rayane.el-sibai@isep.fr

Christine Fricker

Dpartement dinformatique de lENS, cole normale suprieure, CNRS, PSL Research University, 75005 Paris, France
Inria, Paris, France
E-mail: christine.fricker@inria.fr

Abstract: Since its launch, Velib' (the Bike Sharing System -BSS- in Paris) has emerged in the Parisian landscape and has been a model for similar systems in many cities. A major problem with BSS is the stations' heterogeneity caused by the attractivity of some stations located in particular areas. In this paper, we focus on spatial outliers defined as stations having a behavior significantly different from their neighboring stations. First, we propose an improved version of Moran scatterplot to exploit the similarity between neighbors, and we test it on a real dataset issued from Velib' system to identify outliers. Then, we design a new method that globally improves the resources' availability in bike stations by adapting the users' trips to the resources' availability. Results show that with a partial collaboration of the users or a limitation to the rush hours, the proposed method enhances significantly the resources' availability in Velib' system.

Keywords: Outliers detection; spatial data mining; Moran scatterplot; Bike Sharing Systems.

Biographical notes:

Yousra Chabchoub defended her thesis entitled Analysis and modeling of Internet traffic in 2009. Then she joined the Business Intelligence Laboratory (Bilab) at Telecom ParisTech, for a post-doc position, in collaboration with EDF R&D. Since September 2010, she is an associate professor at ISEP. Her research area is data mining, data stream management, and real-time anomalies detection.

Rayane El Sibai received her master degree in software engineering from the Antonine University, Beirut, Lebanon, in 2014. She obtained her Ph.D. degree in Informatique, Télécommunications et Électronique from Université Pierre et Marie Curie - Sorbonne Université, Paris, France, in 2018. Her research interests include Data streams processing, Data summarization, Data quality, and Cloud Computing.

Christine Fricker is a researcher at INRIA Paris. Her research interests concern both theoretical aspects of large stochastic networks and algorithms to manage communication networks. Methods are mainly probabilistic, based on renormalization techniques coming from physical statistics as mean field and fluid limits. Her applications include computer cache-memories, Internet traffic, optical networks, content-centric networks and data center management. She is currently a member of DYOGENE team. After her scolarity at the Ecole Normale Suprieure de Paris, she held her Ph.D. from UPMC (PSL University). She has previously been in MEVAL then RAP team at INRIA. She has publications in Applied Probability, Networking and Internet traffic literature.

1 Introduction

In order to reduce congestion and pollution in cities, Bike Sharing Systems (BSS) have been widely adopted around

the world. The use in these systems is quite simple: a user takes a bike from a given station with a finite number of docks (called capacity of the station) and returns it at the arrival station, near his destination. BSS provides a 24/7 service, complementary to the other public means of transport, with high flexibility and a relatively low cost for the city. But the system is not so easy to manage.

Our work is motivated by the problem of lack of resources, both bikes, and available slots, in bike sharing systems. Indeed, some stations are often almost empty, without enough available bikes or almost full causing users dissatisfaction. In this context, we define outliers as stations with one of the resources lacking while the other stations in the neighborhood are globally balanced. First, we use an adapted version of Moran scatterplot to explore and characterize the neighborhood of such stations. The results show a local heterogeneity in the system: in a small area, bikes availability is often very variable, depending on the station. This local heterogeneity motivates us to propose a new incentive method which encourages users to improve the placement of bikes among the stations. This mechanism is based on a local small change in users trips. In this ecological regulation, users are redirected to another station in the neighborhood of their source or destination to locally reduce the heterogeneity of stations. Even if this problem is addressed for Velib, the BSS of Paris because we have access to data, we are convinced that our study could apply to other BSSs or electric car sharing systems [1] and also [2] for a survey on car sharing systems.

The main tool in this paper is the detection of anomalies in a spatial context, where an object is considered as an outlier if the values of its non-spatial attributes are significantly different from those of other objects in its surrounding. Spatial outliers detection is useful in many applications, such as the detection of abnormal highway traffic patterns [3], the identification of disease outbreaks [4], the detection of tornadoes and hurricanes [5] and the identification of urban soils pollution [6].

Outliers are defined as a set of observations that are inconsistent with the remainder observations. Outliers identification has practical applications in many areas, such as intrusion detection, fraud detection, fault detection and medical informatics [7]. Outliers detection is also an important task in the data analysis process. It aims to detect abnormal patterns and leads to the identification of unusual phenomena, and new knowledge about the monitored environment. To isolate outliers, it is necessary to first characterize the normal observations, which can be provided by the past values of the same object or by the current values issued from other objects in the neighborhood. In this latter case, the outlier is said spatial. In a spatial context, each data is defined with two categories of attributes: spatial attributes and non-spatial attributes. Spatial attributes include the shape, position, and other topological characteristics of the sensor, and they are used to define the neighborhood of the spatial object. Non-spatial attributes include the ID, manufacturer, age, and sensor measure (called behavioral attribute). A spatial outlier represents a local instability and is only compared to the surrounding dataset [8]. This is based on the rule: *"Everything is related to everything else, but nearby things are more related than distant things"* [9].

This paper is organized as follows. We present in Section 2 the existing related works. We describe in Section 3 the dataset for Velib' used in this work and we discuss the problem of allocation of resources in bike sharing systems. We describe in Section 4 the so-called Moran scatterplot technique and propose an original adaptation to Velib' context. We also detail in this section the experiments carried out on Velib' to illustrate the heterogeneity of the system. We present in Section 5 our new solutions to balance BSS and to improve the resource balancing between the stations and validate them on Velib'. Section 6 concludes the paper.

2 Related works

Several algorithms have been developed to detect the outliers in a spatial context. These algorithms can be classified into two categories: graphical-based algorithms, and quantitative-based algorithms. Graphical-based algorithms use visualization. They present, for each spatial point, the distribution of its neighbors and identify outliers as points in specific regions. This category includes Variogram Cloud, Pocket Plot, Scatterplot, and Moran scatterplot methods. Quantitative-based algorithms perform statistical tests to distinguish the outliers from the rest of the data. These methods include Z-Algorithm, Iterative r Algorithm, Iterative z Algorithm, and Median Algorithm.

Scatterplot represents the data in a two-dimensional space where the X-axis represents the values of the non-spatial attribute (the observable) of each object and the Y-axis represents the mean value of the observable of the neighbors of this object. A regression line is used to identify outliers points [10]. Variogram Cloud [11] is a scatterplot between the spatial distance (X-axis) and the difference of the observable values (Y-axis) for each pair of points in the data. Outliers points are identified as pairs of points having a small spatial distance and a big difference for the observables measurements.

The Z statistic approach [12] is one of the most known quantitative-based algorithms for spatial outliers detection. For each spatial object x, S_x denotes the difference between the attribute value of x and the average attribute value of its spatial neighbors. Spatial outliers are simply identified using a threshold based on μ_s and σ_s which respectively represent the mean and the

standard deviation of the attribute value of S over all the spatial objects.

In [13], authors propose two iterative algorithms (iterative r and iterative z) for the detection of spatial outliers. These algorithms detect the outliers on several iterations. Each iteration detects a single outlier and modifies its value in order to reduce its negative impact on its neighbors in the next iteration.

We apply in this paper the spatial outlier detection approach to a particular case study: the performance evaluation of a balancing mechanism in Bike Sharing Systems (BSS). Nowadays, public authorities are more and more encouraging this ecological mean of transportation by expanding the BSS to the suburbs and building new bike paths. Since its launch in 2007, Velib' (the BSS in Paris) has emerged in the Parisian landscape and has been a model for similar systems in many international cities. Velib' provides a significant proportion of people travels as it daily ensures about $110,000$ trips. It involves about 1800 stations with an average distance of 216 meters.

Literature for BSS is now huge. BSS is widely seen as a healthy sustainable urban mobility mode. Indeed many papers present their positive impact on energy consumption and emissions, safety and economy and traffic conditions. See Zhang and Mi [14] and reference therein. A major problem in BSS is the problem of empty stations and full stations caused by the asymmetric attendance to the stations. According to the annual satisfaction survey of Velib', only 50% of users are satisfied with the availability of bikes and docks in the stations [15], [16]. Despite the performed redistribution (moving bikes using trucks), users often find themselves in front of stations that are totally full or empty. The main approach used to study redistribution is optimization for the static one-vehicle problem in Chemla et al. [17] (see also Cruz et al. [18] and references therein) or multi-vehicle problem in Forma et al. [19] and more recently the dynamical problem in Dell'Amico et al. [20]. And very few have been done for designing user incentives for rebalancing strategies. Fricker and Gast [21] proposed explicit 2-choice algorithms, well-known for balancing queueing networks and Haider et al. [22] propose user incentives to decrease redistribution and solve an optimization problem.

In most cities, operators provide open access to real-time status reports on their bike stations. Several studies show the interest of using these data (Froelich *et al.* [23] and Borgnat *et al.* [24], Vogel and Mattfeld [25]). The main objective in the first papers is to understand and characterize the behavior of the users in order to help in designing and planning policy in urban transportation. See a comparative study of the trip data in [26] and docking data [27] in different cities and [24] for classification of the flows of trips in the Lyon bike sharing system, using spatiotemporal characteristics to perform clustering. See also Côme et al. [28] and Bouveyron et al. [29]. But these data are now used for rebalancing purposes. The first paper as far as we know is Gast et al. [30].

The same trends are present for free-floating BSS (FFBSS), recently deployed in many cities. They have to face a more serious imbalance due to less restriction on parking locations. Moreover, rebalancing is more expensive as bikes may be recovered independently. Furthermore, fleet dimensioning and location of the bikes appear as main issues as they can be fixed by optimization in Zhang et al. [31] or deep-learning using data in Pan et al. [32]. And this issue is also studied for docked BSS. See Park and Sohn [33].

3 Dataset description and problem definition

In order to promote innovation and collaboration with scientists, different kinds of data relative to the Velib' system are "Open Data" available for the research community. We performed all the experiments presented in this paper from these datasets.

First, we have static data describing the Velib' stations. They consist of spatial attributes: the geo-coordinates of the station (latitude and longitude), and non-spatial attributes: Id of the station and its capacity (total number of docks). Then, the dynamic data are of two kinds: First, the number of bikes present in each station for each timestamp t are provided in real time. This parameter is varying during the day and is closely dependent on the activity of users. Second, Velib' users' data trips are also available (one file for the trips during a month). A trip is characterized by a departure and arrival timestamp, and a departure and arrival station. The analysis of several months of trips showed a very strong periodicity: trips can be divided into two main categories: the working days and the weekends. Two days of the same category are very similar. We focus in this paper on the working days and we choose to analyze trips during one day: trips that took place on Thursday, October, the 31^{th}, 2013. This duration includes $121,709$ trips, involving 1226 Velib' stations. 1.03% of the trips are related to maintenance (bikes taken for repair) and 1.48% are trips of regulation (bikes moved by trucks).

According to many research studies ([34], [21] and [28]), the Velib' system has some weaknesses caused by the strong attractivity of some stations that can be explained by their location near a railway station or a monument or a business area. Such stations are very often almost empty (no available bike) or completely full (no available dock to put a bike). Despite the performed regulation (bikes moved by trucks), the system is still unbalanced. This unbalanced distribution of bikes among the different stations causes users dissatisfaction. The unbalanced stations are referred to as *problematic* stations. More precisely, we introduce the following

definition: a station is said problematic at timestamp t if its *occupancy ratio* is under 10% or more than 90%.

The *occupancy ratio* of the station, at a timestamp t, is defined as follows:

$$occupancy\ rate\ _t = \frac{Number\ of\ bikes\ present\ at\ t}{Capacity\ of\ the\ station}$$

4 Spatial outliers detection with an improved Moran scatterplot

The objective of this section is to estimate the number of isolated problematic stations at a given timestamp t, which motivates the incentive method detailed in the next section. A good understanding of the current use of the Velib' system and the real needs of the users is mandatory to improve the performance of this system and to plan its future expansion and evolution. An isolated problematic station satisfies both following conditions: First, it is almost empty or almost full at timestamp t. Second, its occupancy ratio is significantly different from the average occupancy of the neighboring stations at the same timestamp t. Thus the isolated problematic stations are among the spatial outliers. In this section, we consider the system at a fixed timestamp t. In order to detect spatial outliers, we opted to use Moran scatterplot [35] that we adapted to the specificities of our context.

4.1 Moran scatterplot

Moran scatterplot [35] illustrates the similarity between an observed value and its neighboring observations. It measures the global spatial autocorrelation over a geographical area, the well-known *Moran's I*. Let us denote by $Z = \{z_i : 1 \leq i \leq n\}$ the set of the different values of the considered observable at a fixed given time t, in n different locations. For each location, the neighborhood is defined based on the geographical distance. Moran scatterplot visualizes the relationship between the values z_i and their neighborhood average $W_i \cdot Z$, where W is a weight matrix that defines a local neighborhood around each location. The observations Z (x-axis) and $W.Z$ (y-axis) are represented by their standardized values.

Moran scatterplot contains four quadrants, corresponding to four types of spatial correlation. The upper-right and lower-left quadrants consist of the locations with positive spatial correlation: association between similar values. In the upper-right quadrant, the high values are surrounded by high neighbors values, while in the lower-left quadrant, the low values are surrounded by low neighbors values. The upper-left and lower-right quadrants incorporate the locations with negative spatial correlation: association between dissimilar values. The upper-left quadrant contains low values surrounded by high neighbors values, while the lower-right quadrant contains high values surrounded by low neighbors values. The objects located in these two quadrants are considered as spatial outliers and can be identified by the statistical test function:

$$Z_i \times \sum_j w_{ij} Z_j < 0$$

W is the contiguity matrix of weights. It indicates the spatial relationship between every couple of objects. W is also called the row-normalized neighborhood matrix. It is based on a threshold d of the geographical distance: i and j are considered as neighbors if and only if $0 \leq d_{ij} \leq d$, where d_{ij} is the distance between i and j. Moreover, all the neighbors of i are equivalent and have the same impact on the calculation of the neighborhood average $W_i \cdot Z$.

Thus, the contiguity matrix W is given by:

$$w_{ij} = \begin{cases} \frac{1}{Number\ of\ neighbors\ of\ i}, & \text{if } 0 \leq d_{ij} \leq d \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

To apply Moran scatterplot to the context of Velib', one has to estimate the crucial parameter d, which represents the highest distance between two neighboring Velib' stations. The choice of d has to achieve the following trade-off: On the one hand, this distance has to be small enough to let the users slightly change their trips at a local scale, and on the other hand, it has to be high to make sure that most stations have a reasonable number of neighboring stations. Velib' stations are generally close to each other and concentrated in the center of Paris and near attractive locations whereas they are distant in the suburbs.

To get in an idea about the geographical proximity between the different stations we plotted, in Figure 1, the distribution of the geographical distance to the nearest station. In Figure 1, the mean distance to the nearest station is of 216 meters. It leads us to propose that the value of parameter d is larger than 300 meters.

To better estimate the choice of d, we plotted in Figure 2 the distribution of the number of neighbors for all the Velib' stations. We tested different values for the threshold distance d (300, 400 and 500 meters). We conclude that a distance of 400 meters is reasonable as, in this case, a given station has on average about 5 neighboring stations. Moreover, with $d = 400$, only 4.4% of the stations do not have any neighboring station.

However, when detecting spatial outliers, the assumption that all the neighbors have the same impact on the neighborhood average may lead to missing some true spatial outliers. In the dataset described in Section 3, there are 1226 stations. As plotted in Figure 3, the capacity of the stations is highly variable between 8 and 114 bikes, with an average of about 31 bikes.

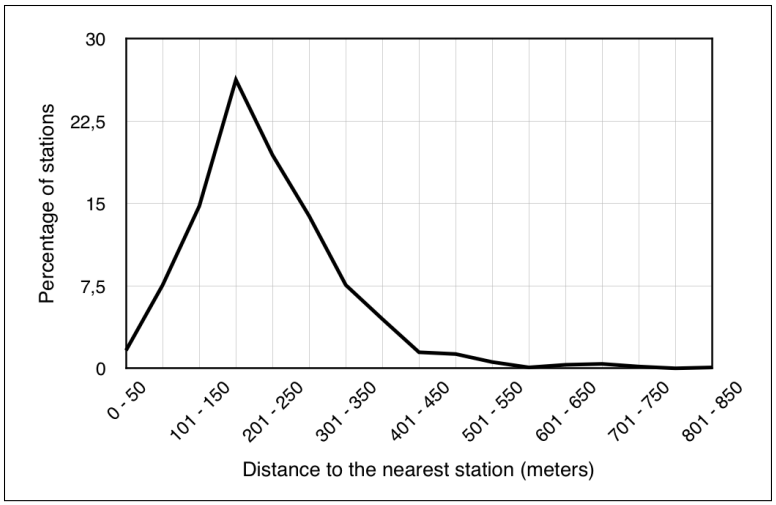

Figure 1: Distribution of the geographical distance to the nearest station in bike sharing system Velib', Paris, France (1226 stations, average distance 215.848 meters).

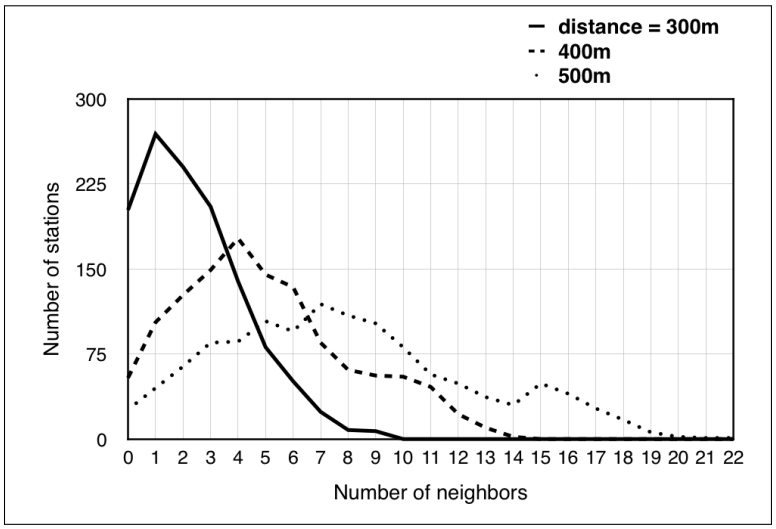

Figure 2: Distribution of the number of *neighbors*, stations at distance less than a threshold distance d, with different values of d: 300, 400 and 500 meters in BSS Velib', Paris, France.

As Velib' stations have different capacities, we defined the occupancy ratio in order to compare normalized bikes availability in these stations. The key idea is that two neighboring stations should have almost the same occupancy ratio if they have similar capacities. That is why the capacity of the station has to be taken into account when calculating the neighborhood occupancy average.

4.2 Improvement of Moran scatterplot using Gower's coefficient

We will replace W with a new weight matrix $\tilde{W}$ also based on the degree of similarity between the station i and the corresponding neighboring stations. This new matrix will take into account the distance and also the difference of capacities between a station and its neighbors. The set N_i of neighbors of station i is defined as previously by the stations with a maximal distance d from station i.

In order to measure the similarity degree between two spatial objects, the Euclidean distance is most often used. However, in our case, the use of this distance is inappropriate since the location and capacity attributes are measured on different scales. Hence, we propose to use the Gower's coefficient [36] to calculate the similarity between two stations. Gower's coefficient is a similarity measure which computes the distance between two instances on each attribute k, and then aggregates all of them to finally calculate the similarity degree.

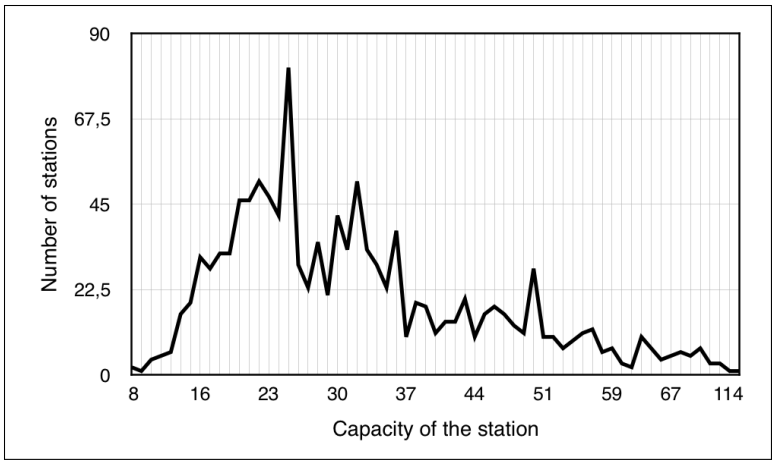

Figure 3: Distribution of station capacity in BSS Velib', Paris, France.

Gower's similarity degree $GOWER_{ij}$ between two stations i and j is defined by:

$$GOWER_{ij} = \frac{\sum_{k=1}^{n} W_{ijk} \times S_{ijk}}{\sum_{k=1}^{n} W_{ijk}} \quad (2)$$

where

- W_{ijk} is the weight associated to the attribute k,
- S_{ijk} is the similarity between two stations i and j for the k^{th} attribute, given by

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}$$

where x_{ik} is the observable attribute k in station i and r_k is a standardization for the attribute k since each attribute is of different unit.

In the context of Velib' stations, we calculate the similarity S_{ij} of the location SD_{ij} and capacity SC_{ij} between two neighboring stations i and j by:

$$SD_{ij} = 1 - \frac{d_{ij}}{d}$$

$$SC_{ij} = 1 - \frac{|Capacity_i - Capacity_j|}{Capacity_{max} - Capacity_{min}}$$

where

- d_{ij} is the distance between the two stations and d is the maximal distance.
- $Capacity_{max}$ and $Capacity_{min}$ are respectively the maximal and minimal stations capacities in the neighborhood of station i.

In this definition, $W_{ijk} = W_{ij}$, where W_{ij} is previously defined by equation (1).

We propose in the following to modify the construction of the contiguity matrix of weights by incorporating the spatial and non-spatial attributes and in a weighted manner in the calculation of the weights associated with neighbors. For each neighboring station j, its new weight $GOWER_{ij}$ regarding the station i is given by equation (2).

The normalization of the contiguity matrix of weights is done per line, so for each station i, the weight of each neighboring station j is divided by the sum of the weights of all the neighboring stations of i.

Thus, the new contiguity matrix $\tilde{W}$ is given by:

$$\tilde{w}_{ij} = \begin{cases} GOWER_{ij}, & \text{if } 0 \leq d_{ij} \leq d \\ 0, & \text{otherwise.} \end{cases}$$

We applied the improved version of Moran scatterplot to detect the isolated problematic stations. Recall that these stations are defined as spatial outliers with a critical occupancy ratio. We used the same data set described in Section 3.

Moran scatterplot representation for the occupancy data of the stations at a fixed timestamp: $10:00$ *am* is given in Figure 4. At this time of day, we can expect that the system is highly unbalanced, as in general in a working day a lot of trips take place in the morning around $8:00$ *am*. The spatial outliers stations (almost 300 stations) are located in the upper-left and lower-right quadrants. One can notice that there are fewer points in these quadrants compared to the locations with positive correlation.

The number of detected isolated problematic stations at $10:00$ *am*, depending on the allowed distance, is given in Table 1. Recall that isolated problematic stations are defined as spatial outliers with critical occupancy ratio. According to this table, there are about 50 isolated problematic stations at $10:00$ *am*. The allowed distance

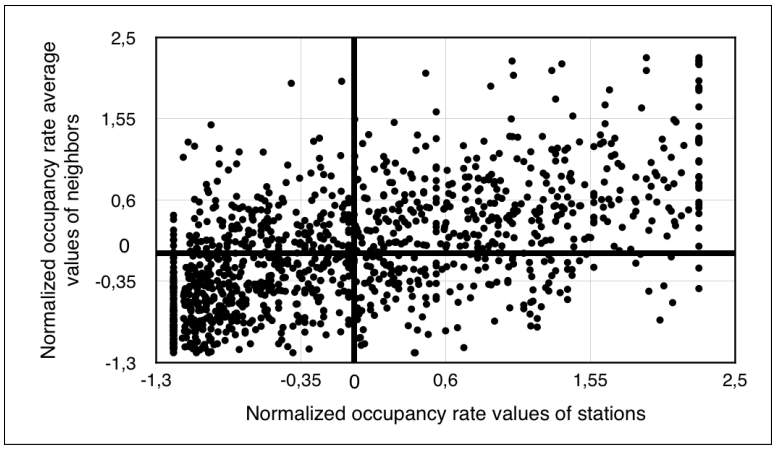

Figure 4: Improved Moran scatterplot based on occupancy data for Velib' bike sharing system, Paris, France, on Thursday 10/31/2013 $10:00$ *am*.

does not have a considerable impact on the number of outliers and the isolated problematic stations. Moreover, with a local change of their trips, Velib' users can enhance the occupancy ratio of about 300 stations (spatial outliers), which represents 24.48% of Velib' stations.

5 Rebalancing the bikes in Velib' system

Our objective is to improve resource availability in the Velib' system by reducing the number of problematic stations. For this purpose, we propose and test in this section a new incentive method, based on a natural and ecological regulation performed by Velib' users themselves. The main idea behind the proposed method is to balance the global system by performing small changes in the trips in small local areas. The preliminary study provided in the previous section proved the presence of isolated problematic stations. Then the aim of this part is to show that around a given problematic station (in a distance smaller than 500 meters), there are many balanced stations (with an occupancy ratio around 50%), which make it possible for Velib' users to balance this problematic station by slightly changing their trips (in practice with an award, extra-time for example).

Using the dataset described in Section 3, we plot in Figure 5 the evolution of the number of current trips during the day (on Thursday 10/31/2013), in order to understand the usage of the Velib' system. One can easily identify two peaks at about $8:00$ am and $6:00$ pm. They clearly correspond to the trips to the offices and the return home after work, as it is a working day.

Users trips unbalance the Velib' system by making some stations problematic (almost empty or almost full). Based on the thresholds of station occupancy introduced before (10% and 90%), the current number of problematic stations is given in Figure 6. Despite the performed bike regulation using tracks, the number of problematic stations during the day remains high. The problematic stations are mainly composed of almost empty stations.

We propose in this section two incentive methods that encourage Velib' users to improve the homogeneity of the stations in terms of occupancy ratio by slightly changing their trips. In the trips dataset, let us denote by A the station where the trip begins and by B where the trip ends. The neighborhood of the station is defined by a distance less than 400 meters.

5.1 First scenario

The key idea is to change the trips as follows: For each trip, in terms of occupancy ratio,

- station A will be replaced by the busiest station in the neighborhood of A,
- station B will be replaced by the emptiest station in the neighborhood of B.

This idea can be easily implemented using a mobile application: The user begins by giving his source and destination stations, respectively A and B. Then the application will propose an alternative itinerary that enables him to save a given amount of time or money. The departure station of this new route is the busiest station in the neighborhood of A and the arrival station is the emptiest station in the neighborhood of B. The new route is of course calculated on-line as it depends on the current state of the system. The new itinerary can be accepted by the user or not. It can be improved with other options for implementation. For example, the user could begin giving his source and destination positions, i.e. his departure and arrival locations, not Velib' stations. Then the application can propose him the best itinerary, choosing among the different stations around

Table 1 Number of detected outliers stations with the improved Moran scatterplot based on occupancy data for Velib' bike sharing system, Paris, France, on Thursday 10/31/2013 10 : 00 *am*

Allowed distance	Outliers	Outliers with critical occupancy ratio
300	297	53
400	334	52
500	339	54

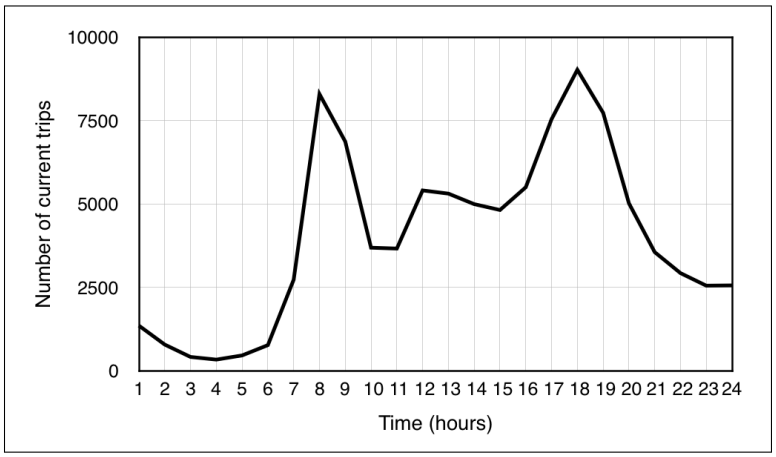

Figure 5: Number of current trips as a function of time during one day time based on trip data for Velib' bike sharing system, Paris, France, on Thursday 10/31/2013.

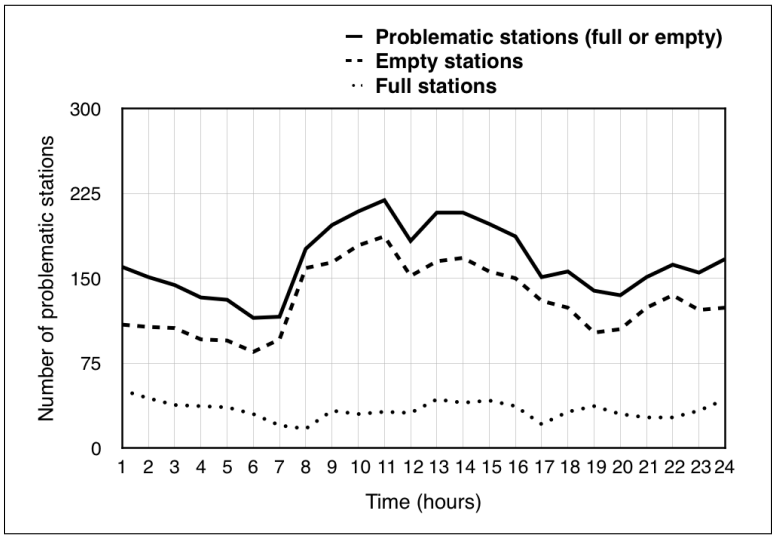

Figure 6: Number of problematic (empty and full), empty and full stations as a function of time during one day, based on occupancy data for Velib' bike sharing system, Paris, France on Thursday 10/31/2013.

these positions. This proposition could be followed or not. The neighborhood is defined by a maximal distance of 400 meters. We checked in section 4 that a value of 400 meters is a suitable distance to have a reasonable number of neighboring stations. Changing the departure station of a trip engenders an extra walk of at most 400 meters to let the user reach the new optimal departure station. Notice that the actual distance is very likely to be greater than the calculated distance because it depends on the chosen path between the departure and the arrival station, whereas the calculated distance only uses the geographical positions of these stations. All the geographical distances used in this paper are Euclidean distances also called Line of Sight (LoS) distances. Given the urban planning in Paris, the real walking or riding distance is approximated by the LoS distance, for small distances. Maybe l_1-distance is more suitable, and one could also derive exactly the stations at a fixed walking distance from a given station. The proposed methodology applies in both cases. To be realistic, the

extra walk engendered by the proposed change has to be compared to the distance of the trip. For this purpose, we plotted in Figure 7 the distribution of trips lengths. Moreover, the trips having the same departure and arrival station are ignored (about 2.27% of the performed trips). The so obtained mean trip length is of 1917 meters, therefore a maximal extra walk of 400 meters can be reasonably proposed. The distribution of trips length can be explained by the fact that the first 30 minutes of the trip are free (45 minutes for the students). The mean duration of the trips is of 14.5 minutes leading to an average speed of 7.93 km/h.

The proposed method is inspired by Velib' + which consists of offering users of Velib' an extra time (that can be cumulated) when they park their bike in a station having a high altitude. The main difference is that *Velib'+* regulation is static: *Velib'+* stations are well known and never change over time, whereas our preferred busiest and emptiest stations dynamically change. They vary during the time depending on their occupancy ratio and the occupancy ratio of their neighboring stations.

5.2 Second scenario

The second scenario consists of performing the same natural regulation proposed in the first scenario, only during the rush hours. The rush hours correspond to the trips to the offices in the morning and the return home at the end of the day. According to Figure 5, these peaks of activities are occurring in the following intervals: [$7h$, $10h$], and [$17h$, $20h$]. They represent 40% of the daily trips. The key idea behind this scenario is to focus on the demands of regular bikers: Changing only the trips of the regular bikers may significantly improve resource availability.

5.3 Experiments and results

Figure 8 presents the impact of the proposed incentive method using the first and the second scenario) on the number of problematic stations. Using the first scenario, The results show a clear decrease in the number of problematic stations throughout the day. The average number of problematic stations drops from 164 in real trips to only 27 by slightly modifying each trip. Starting from a relatively high number of problematic stations (almost 150), users are able to balance almost all these stations within three hours. With the second scenario, which limits the regulation to the rush hours, the number of problematic stations also drops significantly and reaches the performance of the first scenario (permanent regulation) within one hour of the regulation (at 8h and at 18h). Notice that for both scenarios, no new trips are either added or lost. The modification is done with exactly the same number of trips. The real trips are only locally modified. The obtained results confirm our intuition that the global availability of the resources in the Velib' system can be significantly improved by acting locally. This improvement would allow accepting new trips, where originally users are rejected due to a lack of bikes.

The performance of the proposed incentive method can also be measured by the number of spatial outliers in the Velib' system. They consist of stations with an occupancy ratio significantly different from the average occupancy ratio in their neighborhood. These outliers are depicted in Section 4 using Moran scatterplot. The comparison of the number of spatial outliers between the original and modified behaviors (using the first and the second scenario is given in Figure 9. With the improved user behavior, the number of spatial outliers drops significantly, which enhances the balance of the Velib' system.

In Figures 8 and 9, all trips are modified according to the proposed method. It is not a realistic scenario as in real life, many users will not accept to change their departure or arrival station even if they are encouraged by a financial motivation or an extra time offered. To simulate a real-world situation, we plotted in Figure 10 the average number of problematic stations in the day under a variable collaboration rate of the users. One can see that, if only 20% of users accept to change their trips, the number of problematic stations will decrease by half. The decrease in the number of problematic stations is fast (faster than a linear decrease) which is an excellent result as we cannot expect that the majority of users will collaborate.

The number of problematic stations during the day is a good indicator to evaluate the quality of the service offered to Velib' users. However, it cannot entirely qualify service availability. For a given station, the service is considered as interrupted if there is no bike or no dock in this station. In this case, the station is said invalid or out of service. Note that this concerns just one resource: bikes or free docks. To have complete information, we plotted in Figure 11 the average duration of stations invalidity during each one-hour interval of the day, before and after the proposed improvement. One can notice that the mean duration of station invalidity has largely decreased, and likewise, the mean cumulative invalidity duration during the day has been widely improved (cf. Figure 12). According to Figure 12, at the end of the day, the mean cumulative invalidity duration of a Velib' station drops from 141 minutes to only 22 minutes using our first scenario and 68 minutes using the second scenario.

6 Conclusion

We studied in this paper the detection of anomalies in a spatial context. Our use case was the bike sharing system of Paris (Velib'). A Velib' station is considered as a spatial outlier if it is almost empty or almost full while

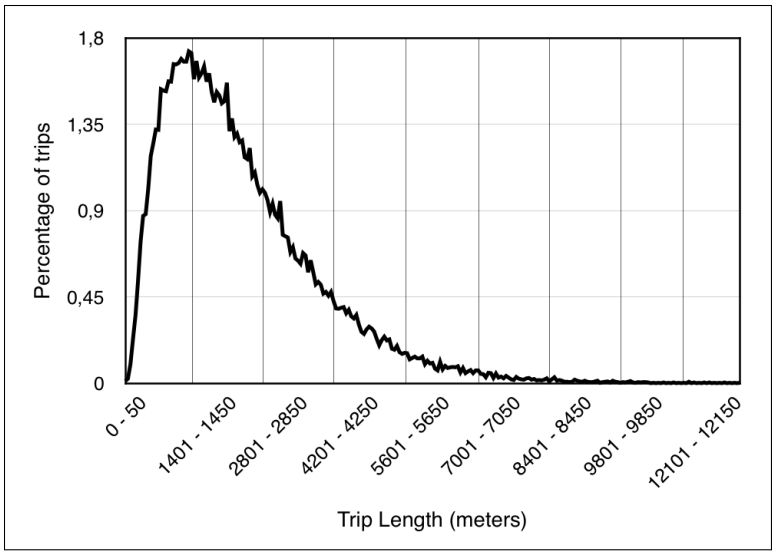

Figure 7: Distribution of the trip length based on trip data for Velib' bike sharing system, Paris, France, on Thursday 10/31/2013.

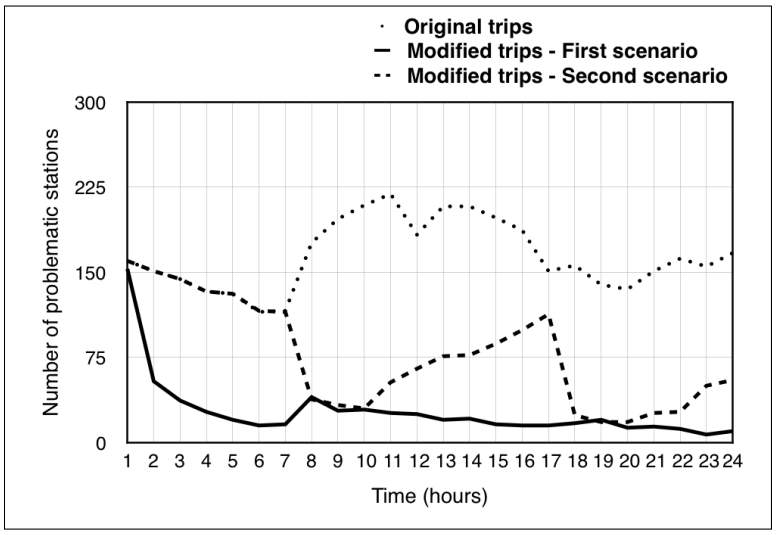

Figure 8: Number of problematic stations as a function of time during one day, based on occupancy data for Velib' bike sharing system, Paris, France on Thursday 10/31/2013.

the surrounding stations are globally balanced. This is due to the problem of heterogeneity in the Velib' system causing resources unavailability and user dissatisfaction. To identify spatial outliers in this context, we used Moran scatterplot. In order to calculate the occupancy distance between two stations, we introduced a similarity weight that takes into account the geographical distance between the stations as well as the difference between their capacities. This degree of similarity is calculated using a robust distance metric called Gower's coefficient. Thereafter, we proposed and tested a new algorithm that locally improves the distribution of resources (bikes and docks) in the stations, and we experimentally validated its efficiency. The results showed that even applied only during the rush hours, the proposed algorithm improves the homogeneity of the Velib' system by reducing the number of outlier stations and the duration of unavailability of the stations during the day, which ultimately leads to the improvement of the user level of satisfaction.

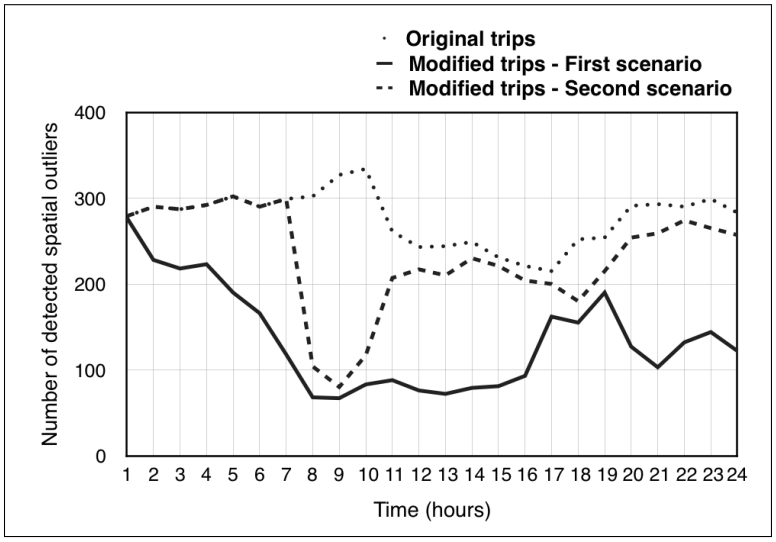

Figure 9: Detected spatial outliers as a function of time during one day, based on occupancy data for Velib' bike sharing system, Paris, France on Thursday 10/31/2013.

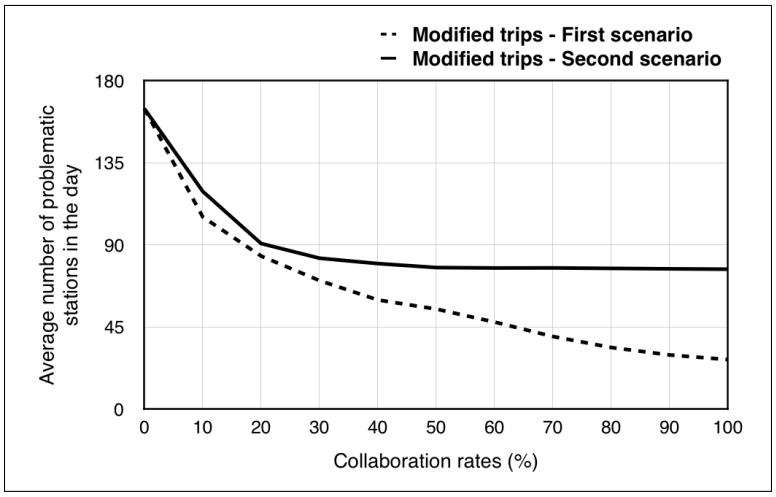

Figure 10: Average number of problematic stations as a function of the user collaboration during one day, based on occupancy data for Velib' bike sharing system, Paris, France on Thursday 10/31/2013.

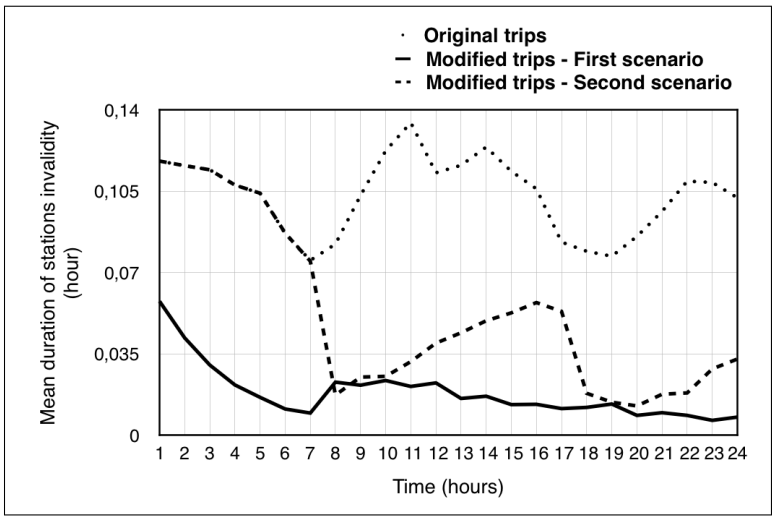

Figure 11: Mean duration of stations invalidity as a function of time during one day, based on occupancy data for Velib' bike sharing system, Paris, France on Thursday 10/31/2013, for the same scenarios as in Figure 10.

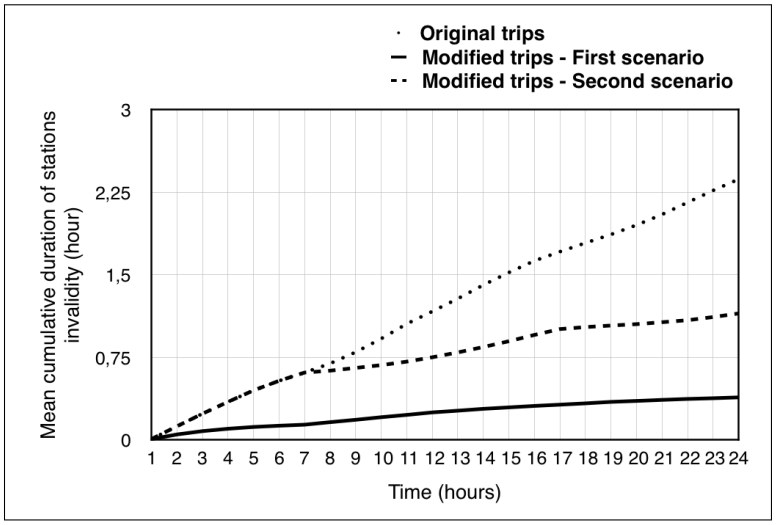

Figure 12: Mean cumulative duration of station invalidity as a function of time during one day, based on both trip and occupancy data for Velib' bike sharing system, Paris, France on Thursday 10/31/2013, for the same scenarios as in Figure 10.

References

[1] Amine Ait-Ouahmed, Didier Josselin, and Fen Zhou. Relocation optimization of electric cars in one-way car-sharing systems: modeling, exact solving and heuristics algorithms. *International Journal of Geographical Information Science*, 32(2):367–398, 2018.

[2] Stefan Illgen and Michael Höck. Literature review of the vehicle relocation problem in one-way car sharing networks. *Transportation Research Part B: Methodological*, 2018.

[3] Shashi Shekhar, Chang-Tien Lu, and Pusheng Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 371–376. ACM, 2001.

[4] Weng-Keen Wong, Andrew Moore, Gregory Cooper, and Michael Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. In *AAAI/IAAI*, pages 217–223, 2002.

[5] Chang-Tien Lu and Lily R Liang. Wavelet fuzzy classification for detecting and tracking region outliers in meteorological data. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pages 258–265. ACM, 2004.

[6] Chaosheng Zhang, Lin Luo, Weilin Xu, and Valerie Ledwith. Use of local moran's i and gis to identify pollution hotspots of pb in urban soils of galway, ireland. *Science of the total environment*, 398(1):212–221, 2008.

[7] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[8] Shashi Shekhar, Michael R Evans, James M Kang, and Pradeep Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.

[9] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[10] Robert Haining. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, 1993.

[11] John Haslett, Ronan Bradley, Peter Craig, Antony Unwin, and Graham Wills. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, 45(3):234–242, 1991.

[12] Shashi Shekhar, Chang-Tien Lu, and Pusheng Zhang. A unified approach to detecting spatial outliers. *GeoInformatica*, 7(2):139–166, 2003.

[13] C-T Lu, Dechang Chen, and Yufeng Kou. Algorithms for spatial outlier detection. In *Third IEEE International Conference on Data Mining*, pages 597–600. IEEE, 2003.

[14] Yongping Zhang and Zhifu Mi. Environmental benefits of bike sharing: A big data-based analysis. *Applied Energy*, 220:296–301, 2018.

[15] Tout sur Vélib. `http://blog.velib.paris.fr/blog/2014/07/15/7-ans-de-velib-des-records-de-frequentation-et-dabonnements/`, 2014.

[16] Rayane El Sibai, Yousra Chabchoub, and Christine Fricker. Using spatial outliers detection to assess balancing mechanisms in bike sharing systems. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 988–995. IEEE, 2018.

[17] Daniel Chemla, Frédéric Meunier, and Roberto Wolfler Calvo. Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization*, 10(2):120–146, 2013.

[18] Fábio Cruz, Anand Subramanian, Bruno P Bruck, and Manuel Iori. A heuristic algorithm for a single vehicle static bike sharing rebalancing problem. *Computers & Operations Research*, 79:19–33, 2017.

[19] Iris A Forma, Tal Raviv, and Michal Tzur. A 3-step math heuristic for the static repositioning problem in bike-sharing systems. *Transportation research part B: methodological*, 71:230–247, 2015.

[20] Mauro DellAmico, Manuel Iori, Stefano Novellani, and Anand Subramanian. The bike sharing rebalancing problem with stochastic demands. *Transportation research part B: methodological*, 118:362–380, 2018.

[21] Christine Fricker and Nicolas Gast. Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity. *Euro journal on transportation and logistics*, 5(3):261–291, 2016.

[22] Zulqarnain Haider, Alexander Nikolaev, Jee Eun Kang, and Changhyun Kwon. Inventory rebalancing through pricing in public bike sharing systems. *European Journal of Operational Research*, 270(1):103–117, 2018.

[23] Jon Froehlich, Joachim Neumann, Nuria Oliver, et al. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI*, volume 9, pages 1420–1426, 2009.

[24] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Céline Robardet, Jean-Baptiste Rouquier, and Eric Fleury. Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03):415–438, 2011.

[25] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523, 2011.

[26] Martin Zaltz Austwick, Oliver OBrien, Emanuele Strano, and Matheus Viana. The structure of spatial networks and communities in bicycle sharing systems. *PloS one*, 8(9):1–17, 2013.

[27] Oliver Obrien, James Cheshire, and Michael Batty. Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 34:262–273, 2014.

[28] Côme Etienne and Oukhellou Latifa. Model-based count series clustering for bike sharing system usage mining: a case study with the vélib'system of paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):39/139/21, 2014.

[29] Charles Bouveyron, Etienne Côme, and Julien Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.

[30] Nicolas Gast, Guillaume Massonnet, Daniël Reijsbergen, and Mirco Tribastone. Probabilistic forecasts of bike-sharing systems for journey planning. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 703–712. ACM, 2015.

[31] Yongping Zhang, Diao Lin, and Zhifu Mi. Electric fence planning for dockless bike-sharing services. *Journal of Cleaner Production*, 206:383–393, 2019.

[32] Ling Pan, Qingpeng Cai, Zhixuan Fang, Pingzhong Tang, and Longbo Huang. A deep reinforcement learning framework for rebalancing dockless bike sharing systems. *arXiv preprint arXiv:1802.04592*, 2018.

[33] Chung Park and So Young Sohn. An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of seoul. *Transportation Research Part A: Policy and Practice*, 105:154–166, 2017.

[34] Yousra Chabchoub and Christine Fricker. Classification of the vélib stations using kmeans, dynamic time wraping and dba averaging method. In *Computational Intelligence for Multimedia Understanding (IWCIM), 2014 International Workshop on*, pages 1–5. IEEE, 2014.

[35] Luc Anselin. *The Moran scatterplot as an ESDA tool to assess local instability in spatial association.* Regional Research Institute, West Virginia University Morgantown, WV, 1993.

[36] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.